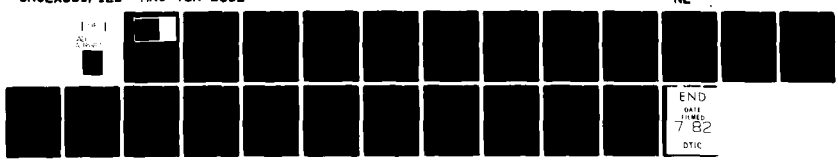
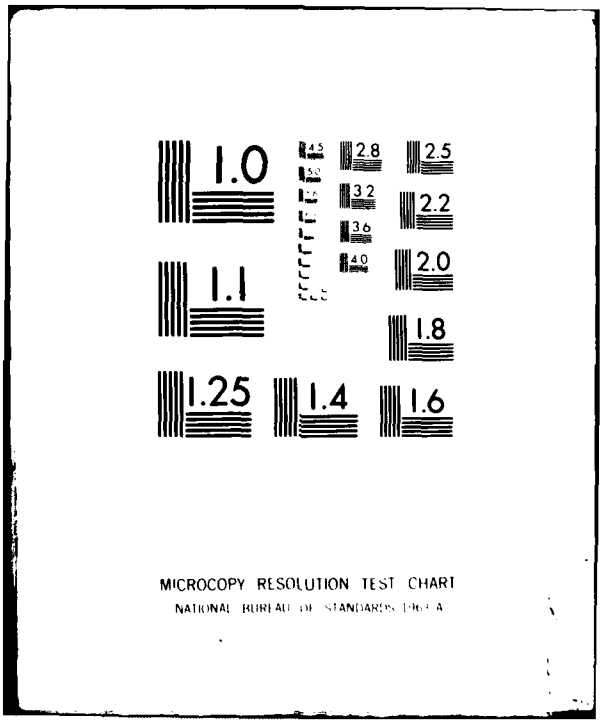
AD-A116 167 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
ROUNDING ERROR IN REGRESSION: THE APPROPRIATENESS OF SHEPPARD'S--ETC(U)
APR 82 A P DEMPSTER, D B RUBIN DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2362 NL



END
DATE FILMED
7 82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

MRC Technical Summary Report #2362

ROUNDING ERROR IN REGRESSION: THE APPROPRIATENESS OF SHEPPARD'S CORRECTIONS

Arthur P. Dempster
and
Donald B. Rubin

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

April 1982

(Received November 9, 1981)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

ROUNDING ERROR IN REGRESSION: THE APPROPRIATENESS OF SHEPPARD'S CORRECTIONS

Arthur P. Dempster and Donald B. Rubin

Technical Summary Report #2362

April 1982

ABSTRACT

We consider three simple approaches to rounding error in least squares regression. The first treats the rounded data as if they were unrounded, the second adds an adjustment to the diagonal of the covariance matrix of the variables, and the third subtracts an adjustment from the diagonal. The third, Sheppard's corrections, can be motivated as maximum likelihood with small rounding error and either (1) joint normal data or (2) normal residuals, "regular" independent variables, and large samples. Although an example and theory suggest that the third approach is usually preferable to the first two, a generally satisfactory attack on rounding error in regression requires the specification of the full distribution of variables, and convenient computational methods for this problem are not currently available.

AMS (MOS) Subject Classification: 62A10, 62F15, 62H12, 62J05

Key Words: EM algorithm, grouping, incomplete data

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

ROUNDING ERROR IN REGRESSION: THE APPROPIATENESS OF SHEPPARD'S CORRECTIONS

Arthur P. Dempster and Donald B. Rubin

1. Introduction

Our purpose is to clarify the problem of adjusting estimated regression coefficients for rounding errors in the data. First, we contrast three methodologies which have been suggested. Two of these lead to simple but different adjustments. The remaining methodology uses likelihood analysis and leads to adjustments which depend on the choice of a prior (marginal) distribution for the design matrix. Second, we derive some details of likelihood analysis for the limiting case of small rounding error. We use our results to point out two circumstances under which likelihood analysis leads approximately to adjustment via Sheppard's (1898) corrections.

Adjustments for rounding error can be surprisingly large, especially when compared to the sampling standard deviation of estimated regression coefficients. In fact, although the standard deviation is generally proportional to $n^{-1/2}$ as sample size n increases, the rounding error adjustment does not decrease as n increases. Furthermore, the size of the adjustment is substantially increased when the design matrix is ill-conditioned, so that well-known numerical accuracy problems associated with ill-conditioned design matrices are complemented by less well-known, but often practically important, rounding error problems.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

An artificial numerical example serves to illustrate potential differences among adjustment techniques. We construct a 4-variate normal distribution of (Y, X_1, X_2, X_3) by specifying zero means and covariance matrix

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ - & 1 & \rho^2 & \rho^2 + \frac{1-\rho^2}{\sqrt{2}} \\ - & - & 1 & \rho^2 \\ - & - & & 1 \end{bmatrix} .$$

This covariance matrix comes from allowing (Y_1, X_1, X_2, X_3) to depend on NID(0,1) variables (Z_1, Z_2, Z_3, Z_4), as follows: $Y = Z_1$, $X_1 = \rho\, Z_1 + \sqrt{1-\rho^2}\, Z_2$, $X_2 = \rho\, Z_1 + \sqrt{1-\rho^2}\, Z_3$, and $X_3 = \rho\, Z_1 + \sqrt{(1-\rho^2)/2}\, (Z_2 + Z_4)$. To obtain a reasonably ill-conditioned design matrix we set $\rho = .9$. Five different regression fits of the form $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$ are summarized in Table 1, along with the associated multiple R^2. Estimated standard deviations are shown in parentheses for two of the fits.

The first column in Table 1 shows the fit obtained with the actual covariance matrix, corresponding to an infinite sample or equivalently to the true model, so that sampling error is zero. The second column of Table 1 is based on a random sample of size $n = 10{,}000$ from the 4-variate normal as defined above with $\rho = .9$. The sampling standard deviations of the estimated b_1, b_2, b_3 are calculated in the usual least squares way. The differences between the first two columns are comfortably within 2σ limits for error.

The remaining three columns of Table 1 are based on analyses of the same sample of 10,000 as the second column, except that the data were rounded before analysis. Specifically, Y was rounded to the form □ • □□□, X_1 was rounded to □•□□, X_2 was rounded to □•□, and X_3 was rounded to □. If we perform least squares fitting on the rounded data, we obtain the results shown in column 3 of Table 1, where the standard deviations are computed from the usual formula ignoring rounding. Comparing columns 1, 2, and 3 shows that rounding can lead to quite degraded accuracy of estimation, and that nominal sampling standard deviations can be misleading indicators of typical error.

Columns 4 and 5 of Table 1 assess the results of two simple adjustment strategies; one gives reasonable results, whereas the other is worse than no adjustment at all. Column 4, labelled Sheppard, is obtained by subtracting a term of the form $\delta^2/12$ from the diagonal elements of sample covariance matrix calculated using the rounded data, where δ denotes the width of the rounding interval, i.e., .001 for Y, .01 for X_1, .1 for X_2, and 1.0 for X_3. Column 5, headed BRB, was obtained in exactly the same way as column 4, except that the appropriate $\delta^2/12$ was added to each diagonal element of the sample covariance matrix. The letters BRB refer to Beaton, Rubin, and Barone (1976) who present an analysis of rounding error in regression which could lead the unwary to an adjustment similar to that shown in column 5. We discuss the BRB analysis in Section 2. Our theoretical results imply that nonnormal but regular distributions for (X_1, X_2, X_3) would produce similar outcomes, where regular is defined in Section 4, but that the outcomes would be different for a nonregular distribution of (X_1, X_2, X_3), uniform for example.

Table 1

	True Model	Unrounded Sample	Uncorrected Rounded	Sheppard Corrected	BRB Corrected
b_1	.2705	.2791 (.0098)	.4450 (.0087)	.2987	.5260
b_2	.2705	.2610 (.0098)	.1634 (.0079)	.2534	.1250
b_3	.4618	.4536 (.0055)	.3738 (.0052)	.4502	.3176
R^2	.9500	.9495	.9393	.9488	.9329

Five sets of regression coeffients and associated multiple correlations and standard deviations.

2. Theoretical Bases for Adjustment.

Probability models for rounding errors must be interpreted with great care if they are to lead to sound adjustments for rounding error. In support of this proposition we review two theoretical arguments which lead to estimates such as those shown in columns 3 and 5 of Table 1. We then introduce likelihood analysis, and draw attention to an essential difference from the other two arguments: likelihood analysis uses the conditional distribution of the unobserved unrounded values given the rounded data, whereas the other arguments use only the marginal distribution of the difference between rounded and unrounded values.

Consider data generated by the familiar linear model

$$\underset{\sim}{Y} = \underset{\sim}{1}\beta_0 + \underset{\sim}{X}\,\underset{\sim}{\beta}_1 + \underset{\sim}{E} \quad . \tag{1}$$

The $n \times 1$ response vector $\underset{\sim}{Y}$ is a linear combination of k predictor variables, where $\underset{\sim}{X}$ denotes the $n \times k$ design matrix giving the values of the k predictors for the n observations, $\underset{\sim}{1}$ is the $n \times 1$ vectors of ones, and $\underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \underset{\sim}{\beta}_1 \end{pmatrix}$ is the $(k+1) \times 1$ vector of linear regression coefficients. The residual variation denoted by the $n \times 1$ random vector $\underset{\sim}{E}$ is assumed to consist of independent $N(0,\sigma^2)$ components. Normality is not required for the first two arguments we present, but complete model specification is needed for likelihood analysis.

In principle, $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ are directly observable whereas $\underset{\sim}{\beta}$ and $\underset{\sim}{E}$ are unknown, but in practice we observe only rounded values $\underset{\sim}{Y}^*$ and $\underset{\sim}{X}^*$ differing from $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ by rounding error which we denote by $\underset{\sim}{e}$ and $\underset{\sim}{d}$, i.e.,

$$(\underset{\sim}{Y}^*, \underset{\sim}{X}^*) = (\underset{\sim}{Y}, \underset{\sim}{X}) + (\underset{\sim}{e}, \underset{\sim}{d}) \quad . \tag{2}$$

After observing $(\underset{\sim}{Y}^*,\underset{\sim}{X}^*)$ we can say with certainty only that the true $(\underset{\sim}{Y},\underset{\sim}{X})$ lies in a rectangular region centered at $(\underset{\sim}{Y}^*,\underset{\sim}{X}^*)$, or equivalently that $(\underset{\sim}{e},\underset{\sim}{d})$ lies in a congruent rectangle translated to the origin.

Theoretical analysis requires hypotheses about the distribution of $(\underset{\sim}{e},\underset{\sim}{d})$. Initially we assume that the rounding error may be regarded as uniformly distributed over the rectangle. In Section 3, we introduce specific notation for the probability density of $(\underset{\sim}{e},\underset{\sim}{d})$.

Combining equations (1) and (2) yields

$$\underset{\sim}{Y}^* = 1\beta_0 + \underset{\sim}{X}^* \underset{\sim}{\beta}_1 + (\underset{\sim}{E}-d\underset{\sim}{\beta}_1+\underset{\sim}{e}) \quad , \tag{3}$$

which has the same form as (1) except the $\underset{\sim}{E}$ is replaced by $\underset{\sim}{E}-d\underset{\sim}{\beta}_1+\underset{\sim}{e}$. Least squares with model (1) can be motivated when the n components of $\underset{\sim}{E}$ are uncorrelated, with zero means and constant variance; consequently, least squares with model (3) can be motivated if a similar condition plausibly holds for the n components of $\underset{\sim}{E}-d\underset{\sim}{\beta}_1+\underset{\sim}{e}$, because then a least squares analysis based on $(\underset{\sim}{Y}^*,\underset{\sim}{X}^*)$ should produce unbiased estimates of $\underset{\sim}{\beta}$ with Gauss-Markoff optimality. Durbin (1954) records the part of this argument depending on zero means of $\underset{\sim}{E}-d\underset{\sim}{\beta}_1+\underset{\sim}{e}$ to conclude that the uncorrected least squares estimate is unbiased. This argument leads to no adjustment for rounding and thus to the estimate in column 3 of Table 1. Cochran (1968) provides further discussion.

Our second theoretical argument is an extension of the Beaton, Rubin, and Barone (1976) study of rounding in regression. BRB use computer simulation to recreate the unknown $(\underset{\sim}{Y},\underset{\sim}{X})$ from the observed $(\underset{\sim}{Y}^*,\underset{\sim}{X}^*)$, and then compute a least squares estimate of $\underset{\sim}{\beta}$ from the simulated $(\underset{\sim}{Y},\underset{\sim}{X})$. Since a single choice of $(\underset{\sim}{Y},\underset{\sim}{X})$ may not be typical, BRB repeat the simulation many times, drawing $(\underset{\sim}{e},\underset{\sim}{d})$ each time from a uniform distribution over the rounding rectangle and computing a least squares $\hat{\underset{\sim}{\beta}}$ for each $(\underset{\sim}{Y},\underset{\sim}{X})$. The main point

of BRB is that the observed variation among these recreated least squares estimates is useful, in the numerical analysis sense, for exhibiting the range of the possible disturbances due to rounding.

BRB illustrate the technique on the much analyzed Longley (1967) data. They average the simulated $\hat{\underset{\sim}{\beta}}$ vectors for the Longley data in order to show a substantial systematic difference between the simulated $\hat{\underset{\sim}{\beta}}$ and uncorrected least squares applied to the rounded data.

The BRB average $\hat{\underset{\sim}{\beta}}$ over a long sequence is approximately

$$\underset{\underset{\sim}{d},\underset{\sim}{e}}{\text{ave}}[(\underset{\sim}{X}^*+\underset{\sim}{d})^T(\underset{\sim}{X}^*+\underset{\sim}{d})]^{-1}[(\underset{\sim}{X}^*+\underset{\sim}{d})^T(\underset{\sim}{Y}^*+\underset{\sim}{e})] \quad . \tag{4}$$

As BRB show, for small rounding intervals, the use of (4) is effectively the same as adding the appropriate $\delta^2/12$ terms to the diagonals of the sample covariance matrix of $(\underset{\sim}{Y}^*,\underset{\sim}{X}^*)$, as illustrated in column 5 of Table 1. With our artificial data, we have an advantage over BRB with the Longley data, because we know the true $\underset{\sim}{\beta}$ and so can see directly that (4) appears to be defective as an adjusted estimator.

We believe that the Durbin-Cochran and BRB approaches fail in our example because the reasoning is insufficiently conditional. An important element in the justification of least squares for the case of unrounded data from model (1) is that, whatever may be the real world processes producing $\underset{\sim}{X}$ and $\underset{\sim}{E}$, the two parts must be unrelated in the sense that knowledge of $\underset{\sim}{X}$ does not provide any information about $\underset{\sim}{E}$. The parallel requirement fails in the case of model (3), because the process of determining $\underset{\sim}{X}$ influences both $\underset{\sim}{X}^*$ and $\underset{\sim}{E}-\underset{\sim}{d}\beta_1+\underset{\sim}{e}$ jointly, i.e., $\underset{\sim}{X}$ determines both $\underset{\sim}{X}^*$ and $\underset{\sim}{d}$, so that $\underset{\sim}{E}-\underset{\sim}{d}\beta_1+\underset{\sim}{e}$ no longer has its initial approximately uniform distribution conditional on the observed $\underset{\sim}{X}^*$.

While the Durbin-Cochran argument implicitly assumes that the a priori distribution of $\underset{\sim}{E}-\underset{\sim}{d}\beta_1+\underset{\sim}{e}$ remains valid given $\underset{\sim}{X}^*$, the BRB argument goes a step further and implicitly assumes the initial distribution of $-\underset{\sim}{d}\beta_1+\underset{\sim}{e}$ holds given both $\underset{\sim}{Y}^*$ and $\underset{\sim}{X}^*$. We say this because the BRB device is to draw from the initial uniform distribution of $\underset{\sim}{d}$ and $\underset{\sim}{e}$ after $\underset{\sim}{Y}^*$ and $\underset{\sim}{X}^*$ are fixed. The implicit assumption fails because observation of $\underset{\sim}{Y}^*$ and $\underset{\sim}{X}^*$ can convey a substantial amount of information about $\underset{\sim}{d}$ and $\underset{\sim}{e}$, especially when large correlations exist among the variables.

The underlying idea of likelihood analysis is to consider the sampling density of $(\underset{\sim}{Y}, \underset{\sim}{X}, \underset{\sim}{Y}^*, \underset{\sim}{X}^*)$ given $\underset{\sim}{\beta}$ and σ^2. Holding $(\underset{\sim}{Y}^*, \underset{\sim}{X}^*)$ fixed at their observed values in this density leads to a function of $(\underset{\sim}{\beta}, \sigma^2, \underset{\sim}{Y}, \underset{\sim}{X})$. To obtain a likelihood function of the parameter $(\underset{\sim}{\beta}, \sigma^2)$, it is necessary to integrate out the random variable $(\underset{\sim}{Y}, \underset{\sim}{X})$ given the fixed $(\underset{\sim}{\beta}, \sigma^2, \underset{\sim}{Y}^*, \underset{\sim}{X}^*)$, which is equivalent to integrating $\underset{\sim}{d}$ and $\underset{\sim}{e}$ over their conditional distribution given $\underset{\sim}{Y}^*$, $\underset{\sim}{X}^*$, $\underset{\sim}{\beta}$ and σ^2.

Likelihood analysis of rounding error was introduced by Fisher (1922) and elaborated by Lindley (1950). Fisher and Lindley show that likelihood analysis justifies the use of Sheppard's corrections when sampling normal populations with small rounding error. In Section 3 we extend the Fisher-Lindley analysis to more general regression models and show that non Gaussian assumptions about the distribution of $\underset{\sim}{X}$ can also lead to likelihood justification for Sheppard's corrections in large samples.

Derivation of Sheppard's corrections from sampling theory may be found in Eisenhart (1947), Haitovsky (1973), Kendall and Stuart (1962), and Wold (1934). Since our experience with simple numerical examples like that shown in Table 1 has led us to mistrust the use of standard sampling theory to justify adjustment for rounding, we do not review literature on sampling bases for Sheppard's corrections.

3. Likelihood Analysis for Small Rounding Error

The analysis here uses the standard linear model (1) with independent $N(0, \sigma^2)$ error components $\underset{\sim}{E}$. As noted in Section 2, likelihood analysis requires that we average over the conditional distribution of $(\underset{\sim}{Y}, \underset{\sim}{X})$ given $(\underset{\sim}{\beta}, \sigma^2, \underset{\sim}{Y}^*, \underset{\sim}{X}^*)$. It follows that distributional assumptions about $\underset{\sim}{X}$ are required to carry out the analysis and that the resulting adjustments may depend on the distribution of $\underset{\sim}{X}$.

In the absence of rounding error, least squares estimates of $\underset{\sim}{\beta}$ can be justified as maximum likelihood estimates based on model (1), independent of any assumed sampling model for $\underset{\sim}{X}$ whose parameter $\underset{\sim}{\theta}$ does not depend on $\underset{\sim}{\beta}$ or σ^2. Our problem is to find corrected maximum likelihood estimates when rounding is present. We restrict details to first order corrections holding in the limit when rounding error is small. Our theoretical analysis is thus directed at finding small adjustments to the least squares estimates such that the adjusted estimates are first order approximations to maximum likelihood estimates.

We suppose that the rows of $\underset{\sim}{X}$ are independently distributed according to a specified model depending on parameter $\underset{\sim}{\theta}$. Denoting the rows of $(\underset{\sim}{Y}, \underset{\sim}{X})$ by $(Y_i, \underset{\sim}{X}_i)$ for $i = 1,2,\ldots,n$, we suppose that $\underset{\sim}{X}_i$ has density $g_i(\cdot|\underset{\sim}{\theta})$. Hence, if we could observe the unrounded $(\underset{\sim}{Y}, \underset{\sim}{X})$ the log likelihood function would be

$$L(\underset{\sim}{\beta}, \sigma^2, \underset{\sim}{\theta}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \underset{\sim}{X}_i\,\underset{\sim}{\beta}_1)^2 + \sum_{i=1}^{n}\log g_i(\underset{\sim}{X}_i|\underset{\sim}{\theta}) \qquad (5)$$

which we call the complete-data log likelihood. We assume throughout our discussion that the rounded data $(\underset{\sim}{Y}^*, \underset{\sim}{X}^*)$ are fixed at their observed values, and hence the unadjusted estimates $\underset{\sim}{\beta}^*$, σ^*, and $\underset{\sim}{\theta}^*$ are also fixed and known, where $(\underset{\sim}{\beta}^*, \sigma^*, \underset{\sim}{\theta}^*)$ is obtained by maximizing (5) after

substituting $(Y^*, \underset{\sim}{X}^*)$ for $(Y, \underset{\sim}{X})$. Note that $\underset{\sim}{\beta}^*$ and σ^* are found by maximizing the first term of (3) whereas $\underset{\sim}{\theta}^*$ is found by maximizing the second term.

Our mathematical discussion is heuristic in the sense that we do not carry out the detailed analysis required to justify our mathematical argument in terms of precise regularity conditions on the functions $g_i(\cdot|\cdot)$ in a neighborhood of $(\underset{\sim}{X}^*, \underset{\sim}{\theta}^*)$. Our mathematical device for obtaining adjustments is as follows. The EM algorithm of Dempster, Laird, and Rubin (1977) applies to the computation of maximum likelihood estimates when data are incomplete, as in our case when $(\underset{\sim}{Y}^*, \underset{\sim}{X}^*)$ is observed but $(\underset{\sim}{Y}, \underset{\sim}{X})$ is not. The EM technique is iterative, but the rate of convergence depends on the fraction of missing information. When the rounding error is vanishingly small, the EM technique converges in one iteration (to the desired first order of approximation), starting from initial estimates $(\underset{\sim}{\beta}^*, \sigma^*, \underset{\sim}{\theta}^*)$.

The required iteration of the EM method has two steps. First, in the E-step, we average the complete-data log-likelihood (5) over the unknown $(\underset{\sim}{Y}, \underset{\sim}{X})$ given the observed rounded $(\underset{\sim}{Y}^*, \underset{\sim}{X}^*)$ and the current estimates $(\underset{\sim}{\beta}^*, \sigma^*, \underset{\sim}{\theta}^*)$. Second, in the M-step, we maximize the resulting function of $(\underset{\sim}{\beta}, \sigma, \underset{\sim}{\theta})$. Since we are concerned with adjusting $\underset{\sim}{\beta}^*$, we need to carry out the E-step only for the first term in (5) which depends only on the familiar sufficient statistics consisting of the sums, sums of squares and products $\sum_{i=1}^{n} (1, Y_i, \underset{\sim}{X}_i)^T (1, Y_i, \underset{\sim}{X}_i)$. Having found the appropriate adjustments to these sufficient statistics, the M-step by definition simply computes the estimates in the usual (i.e., least squares) way from the adjusted statistics. In particular, if we can show the first order corrections to the

quadratic statistics are Sheppard's corrections, then we have shown that least squares applied to Sheppard-corrected basic statistics gives the desired first order corrected maximum likelihood estimates.

To simplify notation, the required details of the E-step are presented here for the sums, sums of squares and products of $\underset{\sim}{Z} = (\underset{\sim}{X}, \underset{\sim}{Y})$. We let $f_i(\cdot|\phi)$ be the density of $\underset{\sim}{Z}_i$ where $\underset{\sim}{\phi} = (\underset{\sim}{\beta}, \sigma, \underset{\sim}{\theta})$.

By expanding $f_i(\underset{\sim}{Z}_i|\phi^*)$ about $\underset{\sim}{Z}_i^*$ we obtain the first term Taylor series approximation

$$f_i(\underset{\sim}{Z}_i, \underset{\sim}{\phi}^*) \doteq f_i^* + \sum_{j=1}^{k} (Z_{ij} - Z_{ij}^*) f_{ij}^* \qquad (6)$$

where Z_{ij} and Z_{ij}^* denote the j^{th} elements in $\underset{\sim}{Z}_i$ and $\underset{\sim}{Z}_i^*$, f_i^* denotes $f_i(\underset{\sim}{Z}_i^*|\underset{\sim}{\phi}^*)$ and

$$f_{ij}^* = \left[\frac{\partial f(\underset{\sim}{Z}_i|\underset{\sim}{\phi}^*)}{\partial Z_{ij}}\right]_{\underset{\sim}{Z}_i=\underset{\sim}{Z}_i^*} . \qquad (7)$$

We suppose that Z_{ij}^* is obtained from Z_{ij} by rounding to the center of an interval of width δ_j, for $j = 1,2,\ldots,k+1$ and $i = 1,2,\ldots,n$. The E-step requires averaging over the conditional distribution of $\underset{\sim}{Z}_i$ given that $\underset{\sim}{Z}_i$ lies in the rounding rectangle centered at $\underset{\sim}{Z}_i^*$. Dividing the marginal density (6) by its integral over the rounding rectangle means, to the desired first order accuracy, dividing (6) by $f_i^* \prod_{j=1}^{k} \delta_j$. Denoting by E the operation of averaging with respect to the appropriately scaled density[1], we find

---

[1] E. g., $E(Z_{ij}-Z_{ij}^*) = \int_{-\frac{\delta_1}{2}}^{+\frac{\delta_1}{2}} \cdots \int_{-\frac{\delta_{k+1}}{2}}^{+\frac{\delta_{k+1}}{2}} t_{ij}[f_i^* + \sum_{j=1}^{k+1} t_{ij} f_{ij}^*] \prod_{j=1}^{k+1} dt_{ij} / f_i^* \prod_{j=1}^{k+1} \delta_j$

where $t_{ij} = (Z_{ij}-Z_{ij}^*)$.

$$E\,(Z_{ij} - Z^*_{ij}) \doteq \frac{\delta_j^2}{12}\frac{f^*_{ij}}{f^*_i} \tag{8}$$

$$E\,(Z_{ij} - Z^*_{ij})^2 \doteq \frac{\delta_j^2}{12}\,, \quad \text{and} \tag{9}$$

$$E\,(Z_{ij} - Z^*_{ij})(Z_{i\ell} - Z^*_{i\ell}) \doteq 0 \tag{10}$$

for all i, j, ℓ with $j \neq \ell$. From (8), (9) and (10) we obtain

$$E(Z_{ij}) = Z^*_{ij} + \frac{\delta_j^2}{12}\frac{f^*_{ij}}{f^*_i} \tag{11}$$

$$E\,(Z_{ij}^2) = Z^{*2}_{ij} + \frac{\delta_j^2}{12}\,(1 + 2Z^*_{ij}\,\frac{f^*_{ij}}{f^*_i}), \tag{12}$$

and

$$E\,(Z_{ij}Z_{i\ell}) = Z^*_{ij}Z^*_{i\ell} + \frac{\delta_j^2}{12}\,Z^*_{i\ell}\,\frac{f^*_{ij}}{f^*_i} + \frac{\delta_\ell^2}{12}\,Z^*_{ij}\,\frac{f^*_{i\ell}}{f^*_i}\,, \tag{13}$$

for all i, j, ℓ with $j \neq \ell$. Note that the ratio f^*_{ij}/f^*_i is the partial derivative of $\log f_i(\underset{\sim}{Z}_i)$ with respect to Z_{ij} at $Z_{ij} = Z^*_{ij}$.

The E-step is completed by summing (11), (12), and (13) over $i = 1,2,\ldots,n$, whence the required adjustments to the sufficient statistics $\frac{1}{n}\sum_{i=1}^{n} Z^*_{ij}$, $\frac{1}{n}\sum_{i=1}^{n} Z^{*2}_{ij}$ and $\frac{1}{n}\sum_{i=1}^{n} Z^*_{ij}Z^*_{i\ell}$, are respectively,

$$\frac{\delta_j^2}{12}\,\frac{1}{n}\sum_{i=1}^{n}\frac{f^*_{ij}}{f^*_i}\,, \tag{14}$$

$$\frac{\delta_j^2}{12}\left[1 + \frac{2}{n}\sum_{i=1}^{n} Z^*_{ij}\,\frac{f^*_{ij}}{f^*_i}\right], \tag{15}$$

and

$$\frac{\delta_j^2}{12}\left[\frac{1}{n}\sum_{i=1}^{n} Z^*_{i\ell}\,\frac{f^*_{ij}}{f^*_i}\right] + \frac{\delta_\ell^2}{12}\left[\frac{1}{n}\sum_{i=1}^{n} Z^*_{ij}\,\frac{f^*_{i\ell}}{f^*_i}\right]; \quad j \neq \ell\,. \tag{16}$$

4. Special Cases leading to the use of Sheppard's Corrections

There are two particular cases where the likelihood analysis of Section 3 for small rounding error leads to Sheppard's corrections for regression coefficients: (1) when the rows of $\underset{\sim}{X}$ are a normal sample and (2) when the rows of $\underset{\sim}{X}$ are a "regular" sample and n tends to infinity. The second case is more fundamental because the entire likelihood analysis leading to the use of maximum likelihood estimates is predicated on large samples.

When $\underset{\sim}{X}$ is normal, $\underset{\sim}{Z}$ is normal with mean say $\underset{\sim}{\mu}$ and variance $\underset{\sim}{\Sigma}$. We start the EM algorithm at the usual moment estimators based on rounded data, say $\bar{\underset{\sim}{z}}^*$ and $\underset{\sim}{S}^*$. Then at $\underset{\sim}{\mu} = \bar{\underset{\sim}{z}}^*$ and $\underset{\sim}{\Sigma} = \underset{\sim}{S}^*$

$$\frac{f_{ij}^*}{f_i^*} = -\frac{1}{2}\frac{\partial}{2\partial z_{ij}}(\underset{\sim}{z}_i - \bar{\underset{\sim}{z}}^*)\underset{\sim}{S}^{*-1}(\underset{\sim}{z}_i - \bar{\underset{\sim}{z}}^*)^T\Big|_{\underset{\sim}{z}_i = \underset{\sim}{z}_i^*} \tag{17}$$

$$= \text{the } j^{\text{th}} \text{ component of } -\underset{\sim}{S}^{*-1}(\underset{\sim}{z}_i^* - \bar{\underset{\sim}{z}}^*) \ .$$

We are now ready to calculate the E-step, that is, to calculate the adjustments to the sufficient statistics. From (14) we see that the adjustment to $\frac{1}{n}\sum_i z_{ij}$ is zero because $\sum_i (\underset{\sim}{z}_i^* - \bar{\underset{\sim}{z}}^*) = \underset{\sim}{0}$. From (15) the adjustment to the quadratic sufficient statistic $\frac{1}{n}\sum_i z_{ij}^2$ is $-\delta_j^2/12$, and from (16) the adjustment for the quadratic sufficient statistic, $\sum_i z_{ij}\, z_{il}$, is zero; these follows because

$$\sum_i \underset{\sim}{z}_i^*(f_{i1}^*/f_i^*, \ldots, f_{ik}^*/f_i^*) = \sum_i \underset{\sim}{z}_i^*[-\underset{\sim}{S}^{*-1}(\underset{\sim}{z}_i^* - \bar{\underset{\sim}{z}}^*)] = -n\underset{\sim}{I} \ . \tag{18}$$

Consequently, when $(\underset{\sim}{Y},\underset{\sim}{X})$ is jointly normal and the rounding errors are vanishingly small, maximum likelihood estimates of regression parameters of $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ are obtained by applying Sheppard's corrections to the covariance matrix of $(\underset{\sim}{Y},\underset{\sim}{X})$: simply subtract $\delta_j^2/12$ from the corresponding diagonal

element of the covariance matrix where δ_j is the width of the rounding rectangle. The likelihood justification for the use of Sheppard's correction with small rounding error and univariate normal data appears in Fisher (1922) and Lindley(1950).

The second case for Sheppard's corrections treats large samples from regular $\underset{\sim}{X}$. As $n \to \infty$, the summations $\frac{1}{n}\sum_{i=1}^{n}$ in (14), (15) and (16) can be replaced by expectations over the distribution of $\underset{\sim}{Z}^*$; doing so, we obtain simplified first order corrections appropriate in large samples. Specifically, the

correction to $\frac{1}{n}\sum_i Z_{ij}$ is

$$\frac{\delta_j^2}{12} E^*\{\frac{\partial}{\partial X_{ij}} [\log f(\underset{\sim}{Z}^*_i|\phi^*)]\} \tag{19}$$

the correction to $\frac{1}{n}\sum_i Z_{ij}^2$ is

$$\frac{\delta_j^2}{12} [1 + 2\, E^*\{Z^*_{ij} \frac{\partial}{\partial Z_{ij}} \log f(\underset{\sim}{Z}^*_i|\underset{\sim}{\phi}^*)\}] \tag{20}$$

and the correction to $\frac{1}{n}\sum_i Z_{ij}Z_{i\ell}$, $j \neq \ell$, is

$$\frac{\delta_j^2}{12} E^*\{Z^*_{i\ell} \frac{\partial}{\partial Z_{ij}} \log f(\underset{\sim}{Z}^*_i|\underset{\sim}{\phi}^*)\} + \frac{\delta_\ell^2}{12} E^*\{\underset{\sim}{Z}^*_{ij} \frac{\partial}{\partial Z_{i\ell}} \log f(\underset{\sim}{Z}^*_i|\underset{\sim}{\phi}^*)\} \tag{21}$$

where ϕ^* is the maximum likelihood estimate of ϕ assuming the rounded data were unrounded, and E^* is the expectation over the distribution of $\underset{\sim}{Z}^*$. As all $\delta_j \to 0$, an expectation over the distribution of rounded data, $\underset{\sim}{Z}^*_i$, will equal the corresponding expectation over the distribution of unrounded data, $\underset{\sim}{Z}_i$, plus terms of order δ_j and higher. Since each expectation in (19) - (21) is multiplied by a factor of δ_j^2, to order δ_j^2, expectations E^* over the distribution of $\underset{\sim}{Z}^*_i$ may be replaced by expectations E over the

distribution of $\underset{\sim}{Z}_i$. Then, if $f(\underset{\sim}{Z} \mid \underset{\sim}{\phi})$ is sufficiently smooth to allow us to interchange the order of integration and differentiation in these expectations, expressions (19) - (21) exactly equal their values under normality, as we now show.

Consider first the expectation in expression (19), $E\{\frac{\partial}{\partial Z_j} [\log f(\underset{\sim}{Z})]\}$, where for notational convenience we suppress the irrevelant subscript i and replace $f(\underset{\sim}{Z}|\underset{\sim}{\phi})$ by $f(\underset{\sim}{Z})$. For all $\underset{\sim}{u}$,

$$\int f(\underset{\sim}{Z}-\underset{\sim}{u})\, d\underset{\sim}{Z} = 1 \quad .$$

Thus,

$$\frac{\partial}{\partial u_j} \int f(\underset{\sim}{Z}-\underset{\sim}{u})\, d\underset{\sim}{Z} = 0 \quad .$$

Letting D^j refer to the partial derivative with respect to the j^{th} argument, and passing the derivative through the integral gives

$$\int [- D^j f(\underset{\sim}{Z}-\underset{\sim}{u})] d\underset{\sim}{Z} = 0 \quad ,$$

and letting $\underset{\sim}{u} = \underset{\sim}{0}$ implies

$$\int \frac{\partial}{\partial Z_j} f(\underset{\sim}{Z})\, d\underset{\sim}{Z} = 0 \quad ,$$

or

$$\int \frac{\partial}{\partial Z_j} [\log f(\underset{\sim}{Z})]\, f(\underset{\sim}{Z})\, d\underset{\sim}{Z} = 0 \quad .$$

Thus,

$$E\{\frac{\partial}{\partial Z_j} [\log f(\underset{\sim}{Z})]\} = 0 \quad .$$

Next consider the expectation in expression (20),

$$E\{Z_j \frac{\partial}{\partial Z_j} [\log f(\underset{\sim}{Z})]\} \quad .$$

For all $\underset{\sim}{u}$,

$$\int Z_j\, f(\underset{\sim}{Z}-\underset{\sim}{u})\, d\underset{\sim}{Z} = u_j + \int Z_j\, f(\underset{\sim}{Z})\, d\underset{\sim}{Z} \quad . \tag{22}$$

Hence

$$\frac{\partial}{\partial u_j} \int Z_j\, f(\underset{\sim}{Z}-\underset{\sim}{u})\, d\underset{\sim}{Z} = 1 \quad .$$

Passing the derivative through the integral gives

$$\int Z_j [-D^j f(\underset{\sim}{Z}-\underset{\sim}{u})] d\underset{\sim}{Z} = 1 \quad .$$

Letting $\underset{\sim}{u} = \underset{\sim}{0}$ gives

$$E\{Z_j \frac{\partial}{\partial Z_j} [\log f(\underset{\sim}{Z})]\} = -1 \quad .$$

Finally, in order to evaluate the expectation in (21), from (22)

$$\frac{\partial}{\partial u_\ell} \int Z_j \, f(\underset{\sim}{Z}-\underset{\sim}{u}) \, d\underset{\sim}{Z} = 0, \qquad \ell \neq j \quad .$$

Passing the derivative through the integral and letting $\underset{\sim}{u} = \underset{\sim}{0}$ gives

$$E\{Z_j \frac{\partial}{\partial Z_\ell} [\log f(\underset{\sim}{Z})]\} = 0, \qquad \ell \neq j \quad .$$

The regularity condition can fail. For example, with uniformly distributed $\underset{\sim}{X}$, the correction to the variance is equal to Sheppard's correction, but opposite in sign, a fact pointed out by Elderton (1938). Moreover, when the second derivative of the density of $\underset{\sim}{X}$ is large in absolute value, as with short, abrupt-tailed distributions, the value of the appropriate maximum likelihood correction with finite n can be quite far from its large sample limit because $\frac{\partial}{\partial X_{ij}} \log f(\underset{\sim}{X}_i|\underset{\sim}{\theta})$ has large variance. Long tailed distributions for $\underset{\sim}{X}_i$ like the Cauchy do not offer any problems with respect to the variance of this partial derivative, but do require larger n for the likelihood of $\underset{\sim}{\theta}$ to be sufficiently concentrated about $\underset{\sim}{\theta}^*$ to justify the approximations used here.

5. Conclusions

With small enough rounding errors and a large enough sample, our analysis and example suggest that Sheppard's corrections applied to the cross products matrix of independent variables will generate appropriate corrections to the regression coefficients in normal linear regression analyses. With moderate rounding errors or moderate sample size, however, it appears that a serious attack on the problem must confront the fact that valid inferences for the regression coefficients will vary with the specification of the distributional form of the independent variables. Further research will be needed before the limits on the practical usefulness of Sheppard's corrections can be stated. Experience with various plausible choices of distributions for the independent variables will require development of feasible computational tools.

REFERENCES

Beaton, A. E., Rubin, D. B. and Barone, J. L. (1976). "The Acceptability of Regression Solutions: Another Look at Computational Accuracy". Journal of the American Statistical Association, 71, (353), pp. 158-168.

Cochran, W. G. (1968). "Errors of Measurement in Statistics". Technometrics, 10, pp. 637-666.

Dempster, A. P., Laird, N. and Rubin, D. B. (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm". The Journal of the Royal Statistical Society - B, 39, 1 pp. 1-38.

Durbin, J. (1954). "Errors in Variables". The Review of the International Statistical Institute, 1, 3, pp. 23-32.

Eisenhart, C. (1947). "Effects of Rounding on Grouping Data". Chapter 4 of Selected Techniques of Statistical Analysis, C. Eisenhart, M. Hastay, W. A. Wallis, pp. 185-223.

Elderton, W. P. (1938). "Correzione dei Momenti Quando la Curva è Simmetrica". Giomale dell' Istituto Italiano degli Attuari, 16, pp. 145-158.

Fisher, R. A. (1922). "On the Mathematical Foundations of Theoretical Statistics". Phil. Trans. Roy. Soc. A, 222, pp. 309-368.

Haitovsky, Y. (1973). Regression Estimation from Grouped Observations. London: Griffin.

Kendall, M. G. and Stuart, A. (1962). The Advanced Theory of Statistics, Volume I. New York: Hafner.

Lindley, D. V. (1950). "Grouping Corrections and Maximum Likelihood Equations". Proceedings of the Cambridge Philosophical Society, 46, pt. 7, pp. 106-110.

Longley, J. W. (1967). "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User". Journal of the American Statistical Association, 62, pp. 819-841.

Sheppard, W. F. (1898). "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equidistant Divisions of a Scale". Proceedings of the London Mathematical Society, 29, pp. 353-380.

Wold, H. (1934). "Sheppard's Corrections Formulae in Several Variables". Skand. Akfuartidskr., 17, 248.

APD/DBR/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2362	2. GOVT ACCESSION NO. AD-A116167	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Rounding Error in Regression: The Appropriateness of Sheppard's Corrections		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Arthur P. Dempster and Donald B. Rubin		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE April 1982 13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

EM algorithm, grouping, incomplete data

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider three simple approaches to rounding error in least squares regression. The first treats the rounded data as if they were unrounded, the second adds an adjustment to the diagonal of the covariance matrix of the variables, and the third subtracts an adjustment from the diagonal. The third, Sheppard's corrections, can be motivated as maximum likelihood with small rounding error and either (1) joint normal data or (2) normal residuals, "regular" independent variables, and large samples. Although an example and theory suggest that the third approach is usually preferable to the first two, a

(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (continued)

generally satisfactory attack on rounding error in regression requires the specification of the full distribution of variables, and convenient computational methods for this problem are not currently available.


DATE
ILME
8